

An Explicit Formula for Likelihood Function for Gaussian Vector Autoregressive Moving-Average Model Conditioned on Initial Observables with Application to Model Calibration.

Du Nguyen
Statistical Alpha Fund Management LLC
du.nguyen@statisticalalpha.com

May 2, 2016

Abstract

We derive an explicit formula for likelihood function for Gaussian VARMA model conditioned on initial observables where the moving-average (MA) coefficients are scalar. For fixed MA coefficients the likelihood function is optimized in the autoregressive variables Φ 's by a closed form formula generalizing regression calculation of the VAR model with the introduction of an inner product defined by MA coefficients. We show the assumption of scalar MA coefficients is not restrictive and this formulation of the VARMA model shares many nice features of VAR and MA model. The gradient and Hessian could be computed analytically. The likelihood function is preserved under the root inversion maps of the MA coefficients. We discuss constraints on the gradient of the likelihood function with moving average unit roots. With the help of FFT the likelihood function could be computed in $O((kp+1)^2T + ckT \log(T))$ time. Numerical calibration is required for the scalar MA variables only. The approach can be generalized to include additional drifts as well as integrated components. We discuss a relationship with the Borodin-Okounkov formula and the case of infinite MA components.

1 Introduction

The main result of this paper is the following:

Theorem 1 *The conditional log-likelihood function of a k -dimension vector autoregressive moving-average model (VARMA)*

$$X_t = \mu + X_{t-1}\Phi_1 + X_{t-2}\Phi_2 + \dots + X_{t-p}\Phi_p + \epsilon_t + \theta_1\epsilon_{t-1} + \dots + \theta_q\epsilon_{t-q} \quad (1)$$

conditioned on the first p observations (X_1, \dots, X_p) of the $T+p$ observations $X_1, \dots, X_p, X_{p+1}, \dots, X_{T+p}$ with $\theta_1, \dots, \theta_q$ are scalars is given by the formula

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \mu, \boldsymbol{\Phi}, \boldsymbol{\Omega}, X_{p+1} \dots X_{T+p} | X_1 \dots X_p) = & -\frac{Tk}{2} \log(2\pi) - \frac{T}{2} \log(\det(\boldsymbol{\Omega})) \\ & - k/2 \log(\det(\lambda' \lambda + I_q)) - \frac{1}{2} \text{Tr}(\mathbf{Z}' \Theta_T^{-1'} K(\boldsymbol{\theta}, T) \Theta_T^{-1} \mathbf{Z} \boldsymbol{\Omega}^{-1}) \end{aligned} \quad (2)$$

where $\theta_0 = 1$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)$, $\boldsymbol{\Phi} = (\Phi_1, \dots, \Phi_p)$, $\boldsymbol{\Omega}$ is the covariance matrix of the i.i.d. Gaussian random variables ϵ_i 's. Here:

$$\mathbf{Z} = \mathbf{X} - \mu - L\mathbf{X}\Phi_1 - \dots - L^p\mathbf{X}\Phi_p$$

$$\mathbf{X} = \begin{pmatrix} X_{p+1} \\ \dots \\ X_{T+p} \end{pmatrix}$$

of size $T \times k$.

$$\begin{aligned} L^i \mathbf{X} &= \begin{pmatrix} X_{p-i+1} \\ \dots \\ X_{T+p-i} \end{pmatrix} \\ \Theta_T &= \begin{pmatrix} \theta_0 & 0 & \dots & 0 & 0 & 0 \\ \theta_1 & \theta_0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{q-1} & \theta_{q-2} & \dots & \dots & 0 & 0 \\ \theta_q & \theta_{q-1} & \theta_{q-2} & \dots & 0 & 0 \\ 0 & \theta_q & \theta_{q-1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \theta_0 \end{pmatrix} \end{aligned} \quad (3)$$

is of size $T \times T$.

$$\lambda = \Theta_T^{-1} \Theta_{*,T-q} \quad (4)$$

is of size $T \times q$. where

$$\Theta_* = \begin{pmatrix} \theta_q & \theta_{q-1} & \cdots & \cdots & \cdots & \theta_1 \\ 0 & \theta_q & \theta_{q-1} & \cdots & \cdots & \theta_2 \\ 0 & 0 & \theta_q & \theta_{q-1} & \cdots & \theta_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & \theta_q \end{pmatrix} \quad (5)$$

is of size $q \times q$ and

$$\Theta_{*,T-q} = \begin{pmatrix} \Theta_* \\ 0_{T-q,k} \end{pmatrix}$$

$$K = K_T K(\boldsymbol{\theta}, T) = I_T - \lambda[\lambda' \lambda + I_q]^{-1} \lambda' = (I_T + \lambda \lambda')^{-1} \quad (6)$$

The optimal value is obtained at

$$\begin{pmatrix} \mu \\ \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_p \end{pmatrix}_{opt} = (\mathbf{X}'_{\theta,LAG} K \mathbf{X}_{\theta,LAG})^{-1} \mathbf{X}'_{\theta,LAG} K \mathbf{X}_{\theta} \quad (7)$$

where:

$$\mathbf{X}_{\theta} = \Theta_T^{-1} \mathbf{X} \quad (8)$$

$$\mathbf{X}_{\theta,LAG} = (\Theta_T^{-1} \mathbf{1} \quad \Theta_T^{-1} \mathbf{X} \quad \Theta_T^{-1} L \mathbf{X} \quad \cdots \quad \Theta_T^{-1} L^p \mathbf{X})$$

and

$$\boldsymbol{\Omega}_{opt}(\theta) = \frac{1}{T} [\mathbf{X}'_{\theta} K \mathbf{X}_{\theta} - \mathbf{X}'_{\theta} K \mathbf{X}_{\theta,LAG} (\mathbf{X}'_{\theta,LAG} K \mathbf{X}_{\theta,LAG})^{-1} \mathbf{X}'_{\theta,LAG} K \mathbf{X}_{\theta}] \quad (9)$$

$\boldsymbol{\Omega}_{opt}$ is positive semi-definite regardless of sample values of X and choice of $\boldsymbol{\theta}$. With these values of Φ_{opt} and $\boldsymbol{\Omega}_{opt}$, (2) is reduced to

$$\begin{aligned} \bar{\mathcal{L}}(\boldsymbol{\theta}, X_{p+1} \cdots X_{T+p} | X_1 \cdots X_p) = & -\frac{Tk}{2} \log(2\pi) - \frac{T}{2} \log(\det(\boldsymbol{\Omega}_{opt}(\theta))) - \\ & \frac{k}{2} \log(\det(\lambda' \lambda + I_q)) - \frac{Tk}{2} \quad (10) \end{aligned}$$

Futher, set

$$\Sigma_T = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q & 0 & \cdots & 0 \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_q & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \gamma_q & \cdots & \gamma_1 & \gamma_0 & \gamma_1 \\ 0 & \cdots & 0 & 0 & \gamma_q & \cdots & \gamma_1 & \gamma_0 \end{pmatrix} \quad (11)$$

with

$$\gamma_l = \begin{cases} (\theta_l + \theta_1\theta_{l+1} + \theta_2\theta_{l+2} \cdots + \theta_{q-l}\theta_q) & \text{for } l = 0, 1, \dots, q \\ 0 & \text{for } l > q \end{cases} \quad (12)$$

Then

$$\Sigma_T^{-1} = \Theta_T^{-1'} K(\boldsymbol{\theta}, T) \Theta_T^{-1} \quad (13)$$

or

$$\Sigma_T = \Theta_T K(\boldsymbol{\theta}, T)^{-1} \Theta_T' \quad (14)$$

also we have

$$\det(\lambda' \lambda + I_q) = \det(\Sigma_T) = \frac{1}{\det(K(\boldsymbol{\theta}, T))} \quad (15)$$

Σ_T is the well-known concentrated covariance matrix in the study of MA(q) process associated with $\boldsymbol{\theta}$ (normalized to standard deviation of noise equals 1). We note this likelihood function is conditional only on the p observations of X , and not on the initial error estimates ϵ in contrast with the typical conditional sum of squares (CSS) approach. In particular, for VMA models with scalar $\boldsymbol{\theta}$, the formula gives an exact likelihood formula. For scalar MA models, the formula for the likelihood function in term of Σ_T is the same as those found in standard text books, e.g. (Box and Jenkins 1970) or (Hamilton 1994). The determinant of Σ_T in (15) is one studied in the strong Szegő limit theorem and the Borodin-Okounkov's determinant formula (Geronimo and Case 1979; Borodin and Okounkov 2000; Basor and H. 2000) in the theory of Toeplitz operators. While we use the Szegő limit theorem to express the large T limit of the determinant in close form, we do

not need to use the Fredholm determinant result in this paper but just mention the context that the determinants in (15) have appeared elsewhere in the literature. The construction of λ and \bar{K} seems new but we could not be sure it has not appeared in the multiple proofs of the Borodin-Okounkov formula. K is related to the matrix A in the second proof of the Borodin-Okounkov's formula in (Basor and H. 2000).

This decomposition permits more effective calculations of the likelihood function when T is large. We note (13), (14), (15) are purely algebraic, depending only on θ and T . To verify them by hand for a few θ and T would be interesting exercises. For example, with $q = 1$ (15) shows the determinant of Σ_T is $1 + \theta_1^2 + \dots + \theta_1^{2T}$, a result well-known in most time series text books.

Using Σ_T , we can rewrite :

$$\begin{pmatrix} \mu \\ \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_p \end{pmatrix}_{opt} = (\mathbf{X}_{\text{LAG}} \Sigma_T^{-1} \mathbf{X}_{\text{LAG}})^{-1} \mathbf{X}'_{\text{LAG}} \Sigma_T^{-1} \mathbf{X} \quad (16)$$

with

$$\mathbf{X}_{\text{LAG}} = \begin{pmatrix} 1 & \mathbf{X} & L\mathbf{X} & \dots & L^p \mathbf{X} \end{pmatrix}$$

$$\Omega_{opt}(\theta) = \frac{1}{T} [\mathbf{X}' \Sigma_T^{-1} \mathbf{X} - \mathbf{X}' \Sigma_T^{-1} \mathbf{X}_{\text{LAG}} (\mathbf{X}'_{\text{LAG}} \Sigma_T^{-1} \mathbf{X}_{\text{LAG}})^{-1} \mathbf{X}'_{\text{LAG}} \Sigma_T^{-1} \mathbf{X}] \quad (17)$$

We will use the notations:

$$\boldsymbol{\theta}(L) = 1 + \theta_1 L + \dots + \theta_q L^q$$

$$\boldsymbol{\Phi}(L) = 1 - \Phi_1 L - \dots - \Phi_p L^q$$

The condition that $\boldsymbol{\theta}$ is scalar is not restrictive, in the sense that given a system with matrix $\Theta(L)$, we can transform it into one with the same transfer function and scalar MA components. The reverse case, scalar Φ is already well-known (for example in Gilbert realization in Linear System literature - for a time series treatment see (Aoki 1987)) - chapter 4.

Let us recall that argument. If $N(L)$ and $D(L)$ are two square matrix polynomials. We can write $T(L) = N(L)^{-1} D(L)$ as a matrix with rational functions entries t_{ij} . Assume all entries $t_{ij}(L) =$

$nt_{ij}(L)/dt_{ij}(L)$ with nt_{ij} and dt_{ij} are relative prime. Take the least common multiple (lcm) polynomial of all the denominators polynomial entries dt_{ij} , call it $d(L)$. $d(L)T(L) = \Phi(L)$ is a polynomial matrix, so we have proved T could be written as $\frac{1}{d(L)}\Phi(L)$ with d scalar and Φ polynomial. Alternatively, and this is what we will use in our simulation result, is to write

$$T(L) = N(L)^{-1}N_A(L)^{-1}N_A(L)D(L) = (\det(N(L)))^{-1}N_A(L)D(L)$$

where $N_A(L)$ is the adjugate matrix of N . The last expression is of the desired form with $\theta = \deg(N(L))$. We note if $\deg(N) \geq \deg(D)$ and $D(L), N(L)$ comes from a minimal realization (via the Kronecker index approach for example) then $\deg(\det(N))$ is the McMillan degree $\delta(T)$ of the process. We will come back to this discussion in the later section on calibration.

The likelihood formula is valid for any sample size, with no restriction on location of roots of θ . However for invertible $\theta(L)$, the terms of Θ_T^{-1} converges as T increase. Here, we apply an observation of (Hansen and Sargent 1980) that we can adjust the transfer function by Blaschke product terms but still preserve the autocovariance function (this is just the trick to replace a root of θ by its inverse.) Further calculations show (Hamilton 1994) that inverting of a root results in multiplying Σ_T with the square of that root. In section 6 we will examine how different components in the above theorem transform under root inverting and verify that the likelihood function above is invariant under the operation of inverting any number of roots. Therefore we can restrict ourselves to working with models with invertible θ only.

There are several advantages in using the above likelihood formula for model calibration. First of all, only q variables need to be optimized numerically, the θ variables. Secondly, we only need to "throw away" only the first p observations. This is in contrast with the CSS method for typical MA models. If the optimization path get to a root of $\theta(L)$ close to 1, coefficients take a long time to decay so a typical CSS needs to throw away many terms before the forecast become stable. Finally, it also compare favorably with the Kalman filter calibration approach. In the multivariate case calibration usually requires Kronecker indices to reduce rank, otherwise the number of variables involved would be $pk^2 + q$. For a process with Kronecker indices $(d_1, \dots, d_i, \dots, d_k)$ with McMillan degree $m = \sum d_i$ the number of

parameters in a typical estimate is (Tsay 1991)

$$m(k+1) + \sum_{j=1}^k [\sum_{i<j} \min\{d_j + 1, d_i\} + \sum_{j>i} \min\{d_j, d_i\}]$$

With our approach, we only have m variables to be optimized numerically while the rest are computed via regression. Also gradients are harder to compute in the traditional approach. Finally we only need to estimate before hand the McMillan degree as an upper bound for q , not the whole set of Kronecker indices. However we advocate further test to reduce the number of non-zero coefficients to simplify the model.

Our approach is hybrid. We use exact likelihood to for MA terms and try to take advantage of the regression formula for AR terms. In effect, it allows for an efficient search for the scalar polynomial $\theta(L)$ that removes the moving average complexity and leave us with AR data where we can apply regression. We can think of this approach as smoothing then regressing, where we have an efficient algorithm to search for smoothing parameters.

We will see in subsequent sections that this conditional likelihood could be computed relatively fast, with the use of convolution algorithm together with Fast Fourier Transform, allowing us to attack very large sample size. Secondly, exact gradient of the likelihood function is computed with ease, resulting in an efficient optimization algorithm. The C++ and R codes developed by the author implement this algorithm. Conceptually, even the evaluation of the Hessian could be done in reasonable time. However for immediate applications the gradient seems sufficient. We note a few algorithms in exact likelihood estimation try to decompose Σ_T to LL' form. In our approach instead of the standard Cholesky decomposition for Σ_T , we use the fact that Σ_T^{-1} could be decomposed to sum of product of matrices that are triangular and Toeplitz (Θ_T^{-1}) or have small number of rows or columns (λ and \bar{K} .) We only need to do Cholesky decomposition of \bar{K} (of size $q \times q$) instead of a matrix of size $T \times T$.

Look at it another way, the theorem says that if $\theta(L)$ is scalar, there exists an inner product defined by a *kernel* given by the positive-definite matrix Σ_T^{-1} . This inner product accounts for the moving average terms. Under this inner product the autoregressive term and likelihood function have simple format similar to the vector autoregressive (VAR) case. The inner product could be evaluated efficiently

with the help of FFT and numerical calibration would need to be done on the θ parameters only.

It is well-known that finite state space linear time invariant systems are exactly those having rational matrix transfer functions. So our result here could be understood as an explicit form of likelihood function for finite state Kalman filter in a MIMO system, conditioned on the first p observations. We do see a possibility that this approach could be useful in calibrating Kalman filters in general.

2 Proof of the theorem

We start out with a general lemma

Lemma 1 *Let A be an arbitrary $T \times q$ matrix. Then*

$$(I_T + AA')^{-1} = I_T - A(I_q + A'A)^{-1}A' \quad (18)$$

In particular the matrix on the right hand side has all eigenvalues in the closed unit disc.

$$\det(I_q + A'A) = \det(I_T - A(I_q + A'A)^{-1}A')^{-1} \quad (19)$$

This is a special case of Woodbury matrix identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

and the Sylvester determinant's entity

$$\det(I_q + AB) = \det(I_T + BA)$$

We note Woodbury matrix identity already has applications in Kalman filter update so we find it interesting but not quite surprising that it plays a core role in our formulation.

Let Z_t be the time series defined by:

$$Z_t = X_t - \mu - X_{t-1}\Phi_1 - \cdots - X_{t-p}\Phi_p = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \quad (20)$$

Assuming we have $n = T + p$ samples $X_1, \dots, X_p, X_{p+1}, \dots, X_{T+p}$ considered as rows of a matrix

$$\hat{\mathbf{X}} = \begin{pmatrix} X_1 \\ \dots \\ X_{T+p} \end{pmatrix}$$

of size $(T + p) \times k$.

Let

$$\mathbf{Z} = \begin{pmatrix} Z_{p+1} \\ \dots \\ Z_{T+p} \end{pmatrix}$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{p+1} \\ \vdots \\ \epsilon_{T+p} \end{pmatrix}$$

$$\boldsymbol{\epsilon}_* = \begin{pmatrix} \epsilon_{p-q+1} \\ \dots \\ \epsilon_p \end{pmatrix}$$

Then the equation (20) gives:

$$\mathbf{Z} = \Theta_T \boldsymbol{\epsilon} + \Theta_{*,T-q} \boldsymbol{\epsilon}_* \quad (21)$$

We note Θ_T^{-1} could be constructed from the power series expansion of $\boldsymbol{\theta}(L)^{-1} = (\theta_0 + \theta_1 L + \dots + \theta_p L^p)^{-1}$ via the Toeplitz map. Recall that for any integer $T > 0$, the map \mathcal{T} mapping a polynomial $(\theta_0 + \theta_1 L + \dots + \theta_p L^p)$ to the matrix Θ_T above preserves addition, unit (1 is mapped to I_T), scalar multiplication and map polynomial multiplication to matrix multiplication. (In algebra language, it is a homomorphism from the matrix algebra of polynomial matrices $\mathbb{R}[L]$ to the algebra $M_T(\mathbb{R})$ of $T \times T$ matrices). Because of this property, Θ_T^{-1} is the image of the truncated power series of $\boldsymbol{\theta}(L)^{-1} = (\theta_0 + \theta_1 L + \dots + \theta_p L^p)^{-1}$ truncated at T terms.

We solve for $\boldsymbol{\epsilon}$ in term of \mathbf{Z} as:

$$\boldsymbol{\epsilon} = \Theta_T^{-1} \mathbf{Z} - \Theta_T^{-1} \Theta_{*,T-q} \boldsymbol{\epsilon}_* \quad (22)$$

Set

$$\lambda = \Theta_T^{-1} \Theta_{*,T-q} \quad (23)$$

We note that the i th-column of λ could be constructed by truncating the first T terms of the power series expansion $(\theta_{q-i} + \theta_{q-i+1} L + \dots + \theta_q L^k) \boldsymbol{\theta}(L)^{-1}$.

Consider the vectorization that sends a $T \times k$ matrix to a $T \times k$ vector, where we expand the rows first:

$$v(A) = \text{vec}(A') \quad (24)$$

Let $\hat{\epsilon}$ be the $T + q$ matrix formed by adding the vector ϵ_* at the beginning of ϵ :

$$\hat{\epsilon} = \begin{pmatrix} \epsilon_{p-q+1} \\ \dots \\ \epsilon_1 \\ \dots \\ \epsilon_{T+p} \end{pmatrix}$$

From the relation

$$(B^T \otimes A) \text{vec}(X) = \text{vec}(AXB)$$

We have

$$v(\Theta \epsilon) = (\Theta \otimes I_k) v(\epsilon)$$

The covariance matrix for $v(\hat{\epsilon})$ is a $(T + q)k \times (T + q)k$ matrix, with diagonal blocks of size $k \times k$ equal to Ω , and zero elsewhere. We will denote it by $\hat{\Omega}_{T+q} = I_{T+q} \otimes \Omega$. The join **pdf** of $\epsilon_{p-q+1}, \dots, \epsilon_{T+p}$ is given by:

$$(2\pi)^{-(T+q)k/2} \det(\hat{\Omega}_{T+q})^{-1/2} \exp\left(-\frac{1}{2} v(\hat{\epsilon})' \hat{\Omega}_{T+q}^{-1} v(\hat{\epsilon})\right) \quad (25)$$

which could be simplified to

$$((2\pi)^k \det(\Omega))^{-(T+q)/2} \exp\left(-\frac{1}{2} [v(\epsilon)' \hat{\Omega}_T^{-1} v(\epsilon) + v(\epsilon_*)' \hat{\Omega}_q^{-1} v(\epsilon_*)]\right)$$

where $\hat{\Omega}_T^{-1}$ and $\hat{\Omega}_q^{-1}$ are diagonal block matrices $I_T \otimes \Omega^{-1}$ and $I_q \otimes \Omega^{-1}$ respectively.

Now we look for the marginal **pdf** with respect to \mathbf{Z} , assuming we ϵ is related to \mathbf{Z} and ϵ_* by equation (22). The approach of taking expectation with respect to initial terms is well-known, where ϵ_* are the initial terms. The **pdf** of \mathbf{Z} is

$$((2\pi)^k \det(\Omega))^{-(T+q)/2} \int_{\epsilon_* \in (\mathbb{R}^k)^q} e^{-\frac{1}{2} [v(\epsilon)' \hat{\Omega}_T^{-1} v(\epsilon) + v(\epsilon_*)' \hat{\Omega}_q^{-1} v(\epsilon_*)]} d\epsilon_*$$

where

$$d\epsilon_* = d\epsilon_{1,1} d\epsilon_{1,2} \dots d\epsilon_{1,k} \dots d\epsilon_{p,1} \dots d\epsilon_{p,k}$$

is the volume component of all the coordinates of ϵ_* .

Expanding using (22), the exponent could be written in the form:

$$\exp(-\frac{1}{2}[v(\epsilon_*)'Av(\epsilon_*) + 2v(\mathbf{Z})'Bv(\epsilon_*) + v(\mathbf{Z})'Cv(\mathbf{Z})])$$

with

$$A = (\lambda' \otimes I_k)(I_T \otimes \Omega^{-1})(\lambda \otimes I_k) + I_q \otimes \Omega^{-1} = (\lambda'\lambda + I_q) \otimes \Omega^{-1} \quad (26)$$

$$B = -(\Theta_T^{-1'} \otimes I_k)(I_T \otimes \Omega^{-1})(\lambda \otimes I_k) = -\Theta_T^{-1'}\lambda \otimes \Omega^{-1} \quad (27)$$

$$C = \Theta_T^{-1'}\Theta_T^{-1} \otimes \Omega^{-1} \quad (28)$$

Here A is a $qk \times qk$ matrix, B is a $Tk \times qk$ matrix and C is a $Tk \times Tk$ matrix. Using the formula

$$\begin{aligned} & \int_{\mathbf{u} \in R^N} \exp(-\frac{1}{2}[\mathbf{u}'A\mathbf{u}^* + 2\mathbf{h}'B\mathbf{u}^* + \mathbf{h}'C\mathbf{h}])d\mathbf{u} \\ &= (2\pi)^{N/2}(\det(A))^{-1/2} \exp(-\frac{1}{2}[\mathbf{h}'(C - BA^{-1}B)\mathbf{h}]) \end{aligned}$$

with $N = qk$ is the dimension of $\mathbf{u} = \epsilon_*$ and

$$\det(A) = \det(\lambda'\lambda + I_q)^k \det(\Omega)^{-q}$$

we deduce:

$$\mathbf{pdf}(\mathbf{Z}) = \frac{\exp(-\frac{1}{2}[v(\mathbf{Z})'[C - BA^{-1}B']v(\mathbf{Z})])}{((2\pi)^k \det(\Omega))^{T/2} \det(\lambda'\lambda + I_q)^{k/2}} \quad (29)$$

The exponent is quadratic in \mathbf{Z} . Note

$$C - BA^{-1}B' = \Theta_T^{-1'}[I_T - \lambda[\lambda'\lambda + I_q]^{-1}\lambda']\Theta_T^{-1} \otimes \Omega^{-1}$$

By lemma 1 $I_T - \lambda[\lambda'\lambda + I_q]^{-1}\lambda'$ is positive definite. The following lemma is well-known in vectorization:

Lemma 2 *For any two matrices of the same size M and N ,*

$$v(M)'v(N) = \text{Tr}(M'N) \quad (30)$$

In particular, if H and Ω are symmetric of size $T \times T$ and $k \times k$ respectively then if X is a matrix of size $T \times k$ we have

$$v(X)'(H \otimes \Omega)v(X) = v(HX)'v(X\Omega) = \text{Tr}(X'HX\Omega)$$

This provides a connection between *vec* and Tr. Set

$$\bar{K} = \bar{K}(\boldsymbol{\theta}) = \lambda' \lambda + I_q$$

Then \bar{K} is an $q \times q$ matrix.

$$K = K(\boldsymbol{\theta}, T) = I_T - \lambda \bar{K}^{-1} \lambda' = I_T - \lambda [\lambda' \lambda + I_q] \lambda'$$

K is a $T \times T$ matrix. Then

$$pdf(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Omega}) = \frac{\exp(-\frac{1}{2} \text{Tr}(\mathbf{Z}' \Theta_T^{-1'} K(\boldsymbol{\theta}, T) \Theta_T^{-1} \mathbf{Z} \boldsymbol{\Omega}^{-1}))}{((2\pi)^k \det(\boldsymbol{\Omega}))^{T/2} \det(\lambda' \lambda + I_q)^{k/2}} \quad (31)$$

From here we have proved (2). Let us now consider the partial optimization problem in $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$ given $\boldsymbol{\theta}$. Substitute

$$\mathbf{Z} = \mathbf{X} - \mu - L\mathbf{X}\Phi_1 - \dots - L^p\mathbf{X}\Phi_p$$

in equation (31), the problem is to find Φ_i minimizing:

$$\begin{aligned} & \text{Tr}((\mathbf{X} - \mu - L\mathbf{X}\Phi_1 - \dots - L^p\mathbf{X}\Phi_p)' \Theta_T^{-1'} K(\boldsymbol{\theta}, T) \Theta_T^{-1} \\ & \quad (\mathbf{X} - \mu - L\mathbf{X}\Phi_1 - \dots - L^p\mathbf{X}\Phi_p) \boldsymbol{\Omega}^{-1}) = \\ & \text{Tr}((\Theta_T^{-1} \mathbf{X} - \Theta_T^{-1} \mu - \Theta_T^{-1} L\mathbf{X}\Phi_1 - \dots - \Theta_T^{-1} L^p\mathbf{X}\Phi_p)' K(\boldsymbol{\theta}, T) \\ & \quad (\Theta_T^{-1} \mathbf{X} - \Theta_T^{-1} \mu - \Theta_T^{-1} L\mathbf{X}\Phi_1 - \dots - \Theta_T^{-1} L^p\mathbf{X}\Phi_p) \boldsymbol{\Omega}^{-1}) \end{aligned} \quad (32)$$

Since the expression is quadratic in μ and Φ_i 's they could be optimized via linear regression with a modified inner product. We form the matrix $\mathbf{X}_{\theta, \text{LAG}}$ of size $T \times (k \times (1 + p))$ as

$$(\Theta_T^{-1} \mathbf{1} | \Theta_T^{-1} \mathbf{X} | \Theta_T^{-1} L\mathbf{X} | \dots | \Theta_T^{-1} L^p \mathbf{X}).$$

If we do not include the constant term μ we could exclude the block $\Theta_T^{-1} \mathbf{1}$. Set

$$\mathbf{X}_\theta = \Theta_T^{-1} \mathbf{X} \quad (33)$$

Then

$$\begin{pmatrix} \mu \\ \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_p \end{pmatrix}_{opt} = (\mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_{\theta, \text{LAG}})^{-1} \mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_\theta$$

is the optimum choice. This is proved in lemma A.1. We note that it does not depend on Ω .

With this choice of Φ_i the minimal value of the quadratic form (32) above is

$$\begin{aligned} & \text{Tr}((\mathbf{X}_\theta)' K \mathbf{X}_\theta \Omega^{-1} - \\ & \quad \text{Tr}((\mathbf{X}_\theta)' K \mathbf{X}_{\theta, \text{LAG}} (\mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_{\theta, \text{LAG}})^{-1} \mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_\theta \Omega^{-1}) \end{aligned} \quad (34)$$

Similar to the maximum likelihood argument for the VAR model, an argument using Jacobian formula relating derivative of det and Tr shows the choice of Ω that minimize the log likelihood is:

$$\mathbf{\Omega}_{opt}(\theta) = \frac{1}{T} (\mathbf{X}_\theta)' K \mathbf{X}_\theta - (\mathbf{X}_\theta)' K \mathbf{X}_{\theta, \text{LAG}} (\mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_{\theta, \text{LAG}})^{-1} \mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_\theta$$

Finally, with that value of $\mathbf{\Omega}$, the matrix inside the trace expression is simply $T I_k$ and hence the trace is Tk . The conditional log-likelihood is:

$$\begin{aligned} \bar{\mathcal{L}}(\theta) = & -\frac{Tk}{2} \log(2\pi) - \frac{T}{2} \log(\det(\mathbf{\Omega}_{opt}(\theta))) \\ & - \frac{k}{2} \log(\det(\lambda' \lambda + I_q)) - \frac{Tk}{2} \end{aligned}$$

We note the formulas appearing here look very much like regular regression/ covariance formulas, but with the inner product is given by $\Theta_T^{-1'} K(\theta, T) \Theta_T^{-1}$.

Let us discuss the relation connecting Σ_T and Θ_T^{-1} and $K(\theta, T)$. This is purely an algebraic equality involving only θ . We observe that the likelihood function, in case of a pure moving average with scalar θ is reduced to the scalar MA(q) model, tensoring with $\mathbf{\Omega}$. Comparing (2) in case $k = 1, p = 0, \mathbf{\Omega} = \sigma^2, \mu = 0$

$$\begin{aligned} \mathcal{L}(\theta, X_1 \cdots X_T) = & -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\det(\mathbf{\Omega})) \\ & - \frac{1}{2} \log(\det(\lambda' \lambda + I_q)) - \frac{1}{2} (\mathbf{X}' \Theta_T^{-1'} K(\theta, T) \Theta_T^{-1} \mathbf{X} \mathbf{\Omega}^{-1}) \end{aligned}$$

with the known formula for MA(q) for example (5.5.5) in (Hamilton 1994) (note Σ_T in our notation is $\sigma^{-2} \mathbf{\Omega}$ in that reference's):

$$\mathcal{L}(\theta) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \log(\det(\Sigma_T)) - \frac{1}{2} (\mathbf{X}' \Sigma_T^{-1} \mathbf{X})$$

we get the required equation.

3 Szegő's limit for MA(q)

For MA(1) it is well known that the large T limit of $\det(\Sigma_T^{-1}) = \det(K_T)$ is just $1 - \theta_1^2$. As this determinant appears in the likelihood function, it would be natural to ask if a similar result hold in general. It turns out that the large T limit of the determinant is always a polynomial.

The strong Szegő's limit theorem ((Szegő 1952; Bingham 2012; Basor and H. 2000)) shows how to compute the large T limit for determinants for truncated Toeplitz matrices arising from certain analytic functions on the unit disc. For Toeplitz matrix generated from rational functions (essentially general VARMA case) the limit is known under state space representation for example in (Gohberg, Kaashoek, and Schagen 1987). (Kramer and Rosenblatt 1993) also mentioned the theorem. For the case MA(q) the expression is very simple but we could not locate a reference so let us state:

Theorem 2 *If $\theta(L) = \prod_{i=0}^q (1 - \lambda_i L)$ is invertible then*

$$\lim_{T \rightarrow \infty} \det(\Sigma_T)^{-1} = \left(\sum_{i=0}^q \theta_i \right) \left(\sum_{i=0}^q (-1)^i \theta_i \right) \prod_{1 \leq i < j \leq q} (1 - \lambda_j \lambda_i)^2 \quad (35)$$

The last term is a symmetric polynomial in λ_i 's so it could also be expressed as a polynomial in θ_i 's

Apply Szegő's limit theorem for the function $a(L) = \theta(L)\theta(L^{-1})$ we have

$$\lim_{T \rightarrow \infty} \det(\Sigma_T) = \exp\left(\sum_{k=1}^{\infty} k(\log(a))_k^2\right)$$

$(\log(a))_k$ means we take the k th coefficients of the Laurent expansion of $\log(a)$. But we have

$$\log(a(L)) = \sum_{i=1}^q \log(1 - \lambda_i L) + \sum_{i=1}^q \log(1 - \lambda_i L^{-1})$$

so we see easily:

$$\log(\theta(L))_k = \sum_{i=1}^q \frac{\lambda_i^k}{k}$$

So the exponent term is

$$\sum_{k=1}^{\infty} \frac{1}{k} \left(\sum_{i=1}^q \lambda_i^k \right)^2 = \sum_{i=1}^q \sum_k \frac{\lambda_i^{2k}}{k} + 2 \sum_{0 \leq i < j \leq q} \sum_k \frac{(\lambda_i \lambda_j)^k}{k}$$

and the limit of Σ_T is

$$\prod (1 - \lambda_i^2)^{-1} \prod (1 - \lambda_i \lambda_j)^{-2} = \prod (1 - \lambda_i)^{-1} \prod (1 + \lambda_i)^{-1} \prod (1 - \lambda_i \lambda_j)^{-2}$$

which is what we have to prove. We can compute easily the expression for $\prod (1 - \lambda_i \lambda_j)$ for small q . The following table summarize $\lim_{T \rightarrow \infty} \Sigma_T^{-1}$ up to $q = 3$.

$q = 1$	$1 - \theta_1^2$
$q = 2$	$(1 - \theta_1^2)(1 - \theta_2)$
$q = 3$	$(1 - \theta_1^2)(1 - \theta_2 + \theta_1 \theta_3 - \theta_3^2)$

For higher q , the polynomial expressed in term of θ 's expands to a large number of monomial terms so it is simpler to evaluate in term of λ_i 's. We will see below they are the same polynomials that enforce invertibility condition for small q . Note that \bar{K}_T is a $q \times q$ matrix so we can also attempt to compute its limit directly. While we could prove in large T limit, entries of \bar{K}_T are rational functions in the θ_i 's, the explicit expressions for the entries are complicated so the fact that the determinant is the reciprocal of a polynomial is interesting.

A similar calculation could also be done for the covariance matrix in the invertible-stable ARMA(p, q) case, the result involves roots of both the numerator and denominator of $\phi(L)/\theta(L)$ (we assume $\phi(0) = \theta(0) = 1$) denoting them by μ_i^{-1} and λ_j^{-1} with the assumption $|\mu_i| < 1, |\lambda_j| < 1 \forall i, j$. The limit formula is:

$$\begin{aligned} \lim_{T \rightarrow \infty} \det(\Sigma_T) &= \prod (1 - \mu_i^2)^{-1} \prod (1 - \mu_i \mu_l)^{-2} \\ &\quad \prod (1 - \lambda_j^2)^{-1} \prod (1 - \lambda_j \lambda_m)^{-2} \prod (1 - \mu_i \lambda_j)^2 \end{aligned} \quad (36)$$

We note the term $\prod (1 - \mu_i \lambda_j)$ is the resultant of $\phi(L)$ and $L^q \theta(L^{-1})$ up to a scale factor.

4 Gradient and Hessian of the likelihood function

Between the two functions $\mathcal{L}(\mu, \theta, \Phi, \Omega)$ and $\bar{\mathcal{L}}(\theta)$, we mostly deal with the second one in calibration. The Hessian of the first one is related standard errors of the regression coefficients. We use the notation ∂_i as short hand for $\frac{\partial}{\partial \theta_i}$. Let $\mathcal{T}(f, T)$ to be the Toeplitz map

mentioned above mapping a power series f to its truncated lower triangular Toeplitz matrix. We have

$$\Theta_T = \mathcal{T}(\boldsymbol{\theta}, T)$$

We have

$$\begin{aligned}\Theta_T^{-1} &= \mathcal{T}(1/\boldsymbol{\theta}, T) \\ \partial_i \Theta_T^{-1} &= -\mathcal{T}(L^i/\boldsymbol{\theta}^2, T)\end{aligned}$$

$$\frac{\partial \Theta_{*,T-q}}{\partial \theta_i} = \begin{pmatrix} 0_{q-i+1, i-1} & I_{q-i+1} \\ 0_{T-q+i-1, i-1} & 0_{T-q+i-1, q-i+1} \end{pmatrix}$$

Here 0_{ab} denotes the zero matrix block of size $a \times b$, so the right hand side of the last equation has an identity matrix of size $q-i+1$ on the right top corner and zero everywhere else.

Put Θ_T^{-1} in block matrix form of sizes $T \times (q-i+1)$ and $T \times (T-q+i-1)$:

$$\Theta_T^{-1} = [(\Theta_T^{-1})_{T, q-i+1} (\Theta_T^{-1})_{T, T-q+i-1}]$$

Then

$$\begin{aligned}\Theta_T^{-1} \frac{\partial \Theta_{*,T-q}}{\partial \theta_i} &= (0_{T, i-1}, \Theta_{T, q-i+1}^{-1}) \\ \partial_i \lambda &= \partial_i [\Theta_T^{-1} \Theta_{*,T-q}] = (\partial_i \Theta_T^{-1}) \Theta_{*,T-q} + (0_{T, i-1}, \Theta_{T, q-i+1}^{-1}) \\ \partial_i \bar{K} &= (\partial_i \lambda)' \lambda + \lambda' \partial_i \lambda \\ \partial_i \bar{K}^{-1} &= -\bar{K}^{-1} (\partial_i \bar{K}) \bar{K}^{-1}\end{aligned}$$

$$\begin{aligned}\partial_i \mathbf{X}_\theta &= -\mathcal{T}(L^i \boldsymbol{\theta}^{-2}, T) \mathbf{X} \\ \partial_i \mathbf{X}_{\theta, \text{LAG}} &= -\mathcal{T}(L^i \boldsymbol{\theta}^{-1}, T) \mathbf{X}_{\theta, \text{LAG}}\end{aligned}$$

For any two matrices A, B depending on θ

$$\begin{aligned}\partial_i AKB &= (\partial_i A)KB + AK(\partial_i B) - A(\partial_i \lambda) \bar{K} \lambda' B - A \lambda \bar{K} (\partial_i \lambda)' B - \\ &\quad A \lambda (\partial_i \bar{K}) \lambda' B \quad (37)\end{aligned}$$

In case $A = B$, the calculation is further simplified as the result is symmetric so we only need to compute half of the terms. Apply (37) with A, B as \mathbf{X}_θ or $\mathbf{X}_{\theta, \text{LAG}}$ we can compute the partial derivatives of

$$D(X, \theta) = \mathbf{X}'_\theta K \mathbf{X}_\theta$$

$$C_{\text{LAG}}(X, \theta) = \mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_{\theta, \text{LAG}}$$

$$B_{\text{LAG}}(X, \theta) = \mathbf{X}'_{\theta, \text{LAG}} K \mathbf{X}_{\theta}$$

$\mathbf{\Omega}_{opt}$ could be written as $D(X, \theta) - B_{\text{LAG}} C_{\text{LAG}}^{-1} B_{\text{LAG}}$ with A, B, C, D are computed from X_{θ} and $X_{\theta, \text{LAG}}$ so we can calculate its derivatives with the help of

$$\partial_i(B' C^{-1} B) = (\partial_i B') C^{-1} B + B C^{-1} (\partial_i B) - B' C^{-1} (\partial_i C) C^{-1} B.$$

Putting everything together we have the partial derivative $\partial_i \mathbf{\Omega}_{opt}$. Furthermore

$$\partial_i \log(\det(\mathbf{\Omega}_{opt})) = \text{Tr}(\mathbf{\Omega}^{-1} \partial_i \mathbf{\Omega}_{opt})$$

$$\partial_i \log(\det(\lambda' \lambda + I_q)) = \text{Tr}(\bar{K}^{-1} \partial_i \bar{K})$$

$$\partial_i \bar{\mathcal{L}}(\boldsymbol{\theta}) = -\frac{T}{2} \partial_i \log(\det(\mathbf{\Omega}_{opt})) - \frac{k}{2} \partial_i \log(\det(\lambda' \lambda + I_q))$$

So we have the gradient of $\bar{\mathcal{L}}$.

Applying the chain rule and various matrix derivative rules we can also compute the Hessian of $\bar{\mathcal{L}}$. Here, we will need

$$\partial_i \partial_j \Theta_T^{-1} = 2 \mathcal{S}(L^{i+j} / \theta^3, T)$$

and derivative of matrix product rules. The calculation is tedious but doable. However we will not pursue its calculation here.

From the general theory of Fisher matrix for maximum likelihood estimates, the Hessian of \mathcal{L} is related to standard error estimates of the $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$. We note the work of Klein and Melard (Klein and Mélard 2014) for general (matrix $\boldsymbol{\theta}$) VARMAX case.

From the expression of \mathcal{L} , the Hessian block $\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mathcal{L})$ is rather complex and could be done through FFT convolution involving convoluting $1/\boldsymbol{\theta}, 1/\boldsymbol{\theta}^2, 1/\boldsymbol{\theta}^3$ with X . We could compute it numerically. We note however the blocks $\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\Phi}}(\mathcal{L})$ and $\mathbf{H}_{\boldsymbol{\Phi}\boldsymbol{\Phi}}(\mathcal{L})$ are rather simple, the first one could be computed from (37) then replacing one of the two \mathbf{Z} terms with \mathbf{X} . The second is a direct generalization of the VAR case:

$$\mathbf{H}_{\boldsymbol{\Phi}\boldsymbol{\Phi}}(\mathcal{L}) = \text{Tr}(\mathbf{X} \Omega_T^{-1'} K \Omega_T^{-1} \mathbf{X} \Omega^{-1})$$

Another approach to gradient would be to consider \mathcal{L} and $\bar{\mathcal{L}}$ as functions of Σ_T , hence as functions γ_i , then express γ_i as functions of $\boldsymbol{\theta}$, as done in (Anderson and Takemura 1986). Their calculations show

the interesting fact that the Jacobian $\frac{\partial \gamma_i}{\partial \theta_j}$ looks closely related to the Szegő limit of the determinant of \tilde{K} :

$$\det\left(\frac{\partial \gamma_i}{\partial \theta_j}\right) = \theta_0^{q+1} \prod (1 - \lambda_i^{-1}) \prod (1 + \lambda_i^{-1}) \prod (1 - \lambda_i^{-1} \lambda_j^{-1})$$

Finally, we could extend this approach for the case where the coefficients of $\boldsymbol{\theta}$ are functions of a finite number of parameters p_j . If we deal with gradients only we only need the matrix $(\frac{\partial \theta_i}{\partial p_j})$ then apply the chain rule. So we can apply our core codes to calibrate even more general models. We will discuss this later in sections on extended models.

5 Computation and Calibration

We have mentioned the computation of $\Omega_T^{-l} S$ is just a convolution of $1/\boldsymbol{\theta}^l$ with S . The only long matrix calculation encountered is of the form $\mathcal{T}(1/f, T)A$ with f is one of $\boldsymbol{\theta}, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3$ and A is one of X or $\Theta_{*, T-q}$. The convolution calculation could be done through Fast Fourier Transform.

First is the calculation of $1/f$. If f is a polynomial of low degree, which we most likely will encounter, we can either use a recursive algorithm to calculate $1/f$ or expand f to partial fractions of form $c/(1 - dL)$ then apply power series expansion to the later. Another method is to use a fast convergent expansion of $1/f$, see for example (Harvey 2011).

To compute the convolution using Fast Fourier transform, we assume coefficients of $1/f$ will be small enough to be ignored after $T_c(\boldsymbol{\theta})$ steps. FFT convolutions algorithms divide T in to short segments where FFT could be computed efficiently, and make use of the fact convolution is transformed to component-wise multiplication after FFT. The segments are then patched together using the overlap-save method, for example.

In practice, since we also need to compute $\Theta_T^{-1} L^i X$ for $i = 0, \dots, p$ it is more convenient to compute $\Theta_{T+p}^{-1} \hat{\mathbf{X}}$ by FFT convolution. As

before

$$\hat{\mathbf{X}} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \\ \vdots \\ X_{T+p} \end{pmatrix}$$

Write $C = \Omega_{T+p}^{-1}$ and $\hat{\mathbf{X}}$ in blocks of size $p - i$, T and i :

$$\Theta_{T+p}^{-1} \hat{\mathbf{X}} = \begin{pmatrix} C_{[1:(p-i), 1:(p-i)]} & 0 & 0 \\ C_{[(p-i+1):(T+p-i), 1:(p-i)]} & \Theta_T^{-1} & 0 \\ C_{[(T+p-i+1):, 1:(p-i)]} & C_* & C_{**} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{X}}_{[1:(p-i)]} \\ L^i \mathbf{X} \\ \hat{\mathbf{X}}_{[(T+p-i+1):]} \end{pmatrix}$$

We see the block of rows $p - i + 1$ to $T + p - i$ of $\Theta_{T+p}^{-1} \hat{\mathbf{X}}$ is

$$C_{[(p-i+1):(T+p-i), 1:(p-i)]} \hat{\mathbf{X}}_{[1:(p-i)]} + \Theta_T^{-1} L^i \hat{\mathbf{X}}$$

From here $\Theta_T^{-1} L^i \mathbf{X}$ is backed out by subtracting these rows by the first term. The submatrix of Θ_{T+p} could be expressed in term of segments of the power series expansion of $\boldsymbol{\theta}(L)^{-1}$. This adjustment computation is at cost of $O(T \times i)$ each for a total cost of $O(Tp(p+1)/2)$ total. For large T the contribution of the adjustment block decays relatively fast.

We note the inversion of $\boldsymbol{\theta}$ is $O(T \log(T))$ and the convolution is $O(kT \log(T))$ if we use FFT. Linear regression is $O((kp+1)^2 T)$ if T is much larger than k . Overall, the computation of the likelihood function is of order $O(c_1(kp+1)^2 T + c_2 k T \log(T))$ for constants c_1, c_2 . This is already an improvement over the $O(T^2)$ estimate for standard Kalman filter calculation. See, however (Pnevmatikakis et al. 2014) for an approximation of time $O(T + \log(T))$.

The strength of the method is in calibration. We only need to optimize the function $\mathcal{L}(\boldsymbol{\theta})$ in the $\boldsymbol{\theta}$ parameter. This function is much simpler to compute and optimize than the traditional Kalman filter approach. First of all, the function is symmetric, we can save half of the calculation by applying transposes. The matrices encountered here are triangular Toeplitz matrices. Secondly, K is the only large square matrix encountered. But we never need to calculate K directly. Recall

$$\bar{K} = I_q + \lambda' \lambda$$

\bar{K} is of size q , and also symmetric. Therefore we can do a Cholesky decomposition

$$\bar{K} = C_K C_K'$$

Here C_K is a lower triangular matrix of size $q \times q$. so all the subsequent calculation are all based on triangular matrices. Since

$$K = I_T - \lambda \bar{K}^{-1} \lambda'$$

to compute $N'KM$, with N and M has a small number of columns, and of T rows, we actually need to compute $N'M$, $N'\lambda C_K^{-1'}$ and $C_K^{-1}\lambda'M$. All the matrix multiplications and inversions here involve matrices of size $m \times T$ with $m \ll T$.

The next question is how to determine p and q . Here we come back to the earlier discussion on expressing $N(L)^{-1}D(L)$ to the scalar denominator form, with $\theta(L) = \det(N(L))$ and $\Phi(L) = N_A(L)D(L)$. IF N and D comes from a minimal realization, we see both $\deg(\det(N(L)))$ and $\deg(N_A(L))D(L)$ are smaller than the McMillan degree m . So we can choose to maximize likelihood with both p and q equals to the McMillan degree, then apply tests to determine which higher order terms could be eliminated, with help of information criteria. This will need further research since this suggestion may be far from optimal if the actual degree of q is much smaller than m . We expect canonical correlation analysis to play a role here.

Let us now discuss the actual maximization of the likelihood function. We recall again that the likelihood formula is valid for θ with roots anywhere in the complex plane. However, if we try to evaluate it with θ with at least one root inside the unit disc, both Θ_T and Ω will assume a very large value, even though the likelihood function remains finite. The root inversion maps, discussed in section (6) give us in every case a pair $(\theta_{IR}, \Omega_{IR})$ with the same likelihood function value, where θ_{IR} has roots outside the unit disc. The case of roots on the unit circle is special. We will discuss it briefly in section 6.

We can always choose the initial model to be invertible, and we must if we want the system to be identifiable. The region of θ 's where $\theta(L)$ has roots outside of the unit disc is a convex connected region, as discussed in the section 6. The likelihood function is not in general convex. Indeed, a careful analysis of scalar MA(1) case in (Davis and Dunsmuir 1996) shows it could have several local maxima with different asymptotics. In the next section we discuss some properties of this region, and how to choose the initial points for the calibration.

For the actual calibration the reader can choose his favorite gradient-based optimizer, L-BFGS-B is the author's method of choice for this case. We force the optimization to stay within the invertibility region by assigning a large value to the cost function when it is outside. Presumably there may be a better algorithm taking to account the shape of the region as well as the way the gradient transforms under the root inversion map.

6 Root inversion maps

Hansen and Sargent (Hansen and Sargent 1980) (also (Hamilton 1994) for a detailed exposition) proposed a scheme to transform any MA systems to one with roots within the unit disc. Under that scheme the autocovariance-generating function of the new model is the same as the original one. We show here that these transformations works in the case of VARMA with scalar θ and check that they also preserve the conditional log-likelihood. This is a direct generalization of the similar result in the scalar moving average case.

Assuming the equation $\theta(L) = 0$ having roots $\lambda_1^{-1}, \dots, \lambda_q^{-1}$:

$$\theta(L) = \prod_{l=1}^q (1 - \lambda_l L)$$

Recall $Z_t = \sum_{i=0}^q \theta_i \epsilon_{t-i}$, so Z is a VMA process. Let γ_i be given in (12). Consider the autocovariance-generating function of Z :

$$g(z; \theta; \Omega) = \left(\sum_{i=-\infty}^{\infty} \gamma_i z^i \right) \Omega = \prod_{l=1}^q (1 - \lambda_l z) (1 - \lambda_l z^{-1}) \Omega$$

Let us recall how the root inversion maps are constructed. Choose a subset of indices i_1, \dots, i_r corresponding to $\lambda_{i_1} \cdots \lambda_{i_r}$ and consider the polynomial

$$\theta_{IR}(L) = \theta_{IR_{i_1, \dots, i_r}}(L) = \prod_{i \notin \{i_1 \cdots i_r\}} (1 - \lambda_i L) \prod_{i \in \{i_1 \cdots i_r\}} (1 - \lambda_i^{-1} L) \quad (38)$$

and the covariance matrix

$$\Omega_{IR} = (\lambda_{i_1} \cdots \lambda_{i_r})^2 \Omega \quad (39)$$

The following theorem summarizes some important the properties of this map:

Theorem 3 *Under the root inversion map corresponding to $i_1 \cdots i_r$, the autocovariance generating function of $\boldsymbol{\theta}_{IR}$ is invariant:*

$$g(z; \boldsymbol{\theta}_{IR}; \boldsymbol{\Omega}_{IR}) = g(z; \boldsymbol{\theta}; \boldsymbol{\Omega}) \quad (40)$$

Let $\Theta_{IR;T}$ and $\bar{K}(\boldsymbol{\theta}_{IR})$, $K(\boldsymbol{\theta}_{IR}, T)$ be the Toeplitz matrix and K -matrix corresponding to $\boldsymbol{\theta}_{IR}$ we have:

$$\Sigma_{IR;T} := \Theta_{IR} K_{IR}^{-1} \Theta'_{IR} = (\lambda_{i_1} \cdots \lambda_{i_r})^{-2} \Sigma_T \quad (41)$$

$$\det(\bar{K}(\boldsymbol{\theta}_{IR})) = (\lambda_{i_1} \cdots \lambda_{i_r})^{-2T} \det(\bar{K}(\boldsymbol{\theta})) \quad (42)$$

$$\begin{pmatrix} \mu \\ \Phi \end{pmatrix}_{opt}(\boldsymbol{\theta}_{IR}) = \begin{pmatrix} \mu \\ \Phi \end{pmatrix}_{opt}(\boldsymbol{\theta}) \quad (43)$$

$$\boldsymbol{\Omega}_{opt}(\boldsymbol{\theta}_{IR}) = (\lambda_{i_1} \cdots \lambda_{i_r})^2 \boldsymbol{\Omega}_{opt}(\boldsymbol{\theta}) \quad (44)$$

The conditional likelihood functions are also invariant under this transformation.

$$\mathcal{L}(\boldsymbol{\theta}_{IR}, \mu, \boldsymbol{\Phi}, \boldsymbol{\Omega}_{IR}) = \mathcal{L}(\boldsymbol{\theta}, \mu, \boldsymbol{\Phi}, \boldsymbol{\Omega}) \quad (45)$$

$$\bar{\mathcal{L}}(\boldsymbol{\theta}_{IR}) = \bar{\mathcal{L}}(\boldsymbol{\theta}) \quad (46)$$

Let $J[IR]$ be the gradient of IR . The gradient of $\bar{\mathcal{L}}$ transforms under:

$$\nabla \bar{\mathcal{L}}(\boldsymbol{\theta}) = \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}_{IR}) J[IR]$$

The proof of (40), (41) are the same as the scalar case in (Hamilton 1994). (43) is straight forward from its expression. (42) follows from $\det(\bar{K}) = \det(\Sigma_T)$. The last two equations come from direct substitutions.

Note $\lambda_1, \dots, \lambda_q$ are roots of $L^q \boldsymbol{\theta}(L^{-1})$, we may sometime call them roots unless there is confusion. The invertibility condition is $|\lambda_i| \leq 1$.

We have the Vieta map \mathbf{v} from roots $(\lambda_1, \dots, \lambda_q)$ to coefficients $(\theta_1, \dots, \theta_q)$. This map is well defined and algebraic, given by symmetric polynomial equations:

$$\theta_l = \sum_{i_1, \dots, i_l} (-1)^l \prod_{i \in i_1, \dots, i_l} \lambda_i$$

The map \mathbf{v} is not one-to-one, at least any permutation of $\lambda_1, \dots, \lambda_q$ give us the same coefficients. Now the root inversion maps described above are defined on the root space, but only defined on the coefficient

space after we have chosen a partial inverse of \mathbf{v} , which is a particular ordering of the roots. With this in mind we will determine the effect of root inversion on the gradient of the likelihood function. We will use the chain rule and the implicit function theorem and for that we will need the Jacobian of \mathbf{v}

$$J_{\mathbf{v}} = \left(\frac{\partial \theta_i}{\partial \lambda_j} \right)$$

$$\frac{\partial \theta_i}{\partial \lambda_j} = (-1)^{i-1} \sum_{(\mathbf{l}): |\mathbf{l}|=i-1} \prod_{j \notin \mathbf{l}} \lambda_{l \in (\mathbf{l})}$$

Which denote sum over products of $i - 1$ elements that does not contain j . So the j th column is just the coefficients of the expansion of $-\boldsymbol{\theta}(L)/(1 - \lambda_j L)$. A trick we use repeatedly is to evaluate complex expressions at roots of unity, then apply IDFT to compute the coefficients. In the code we use this trick to evaluate the later expression. We note that the Jacobian could be complex if some of the roots are complex. Also $J_{\mathbf{v}}$ is not invertible at roots with multiplicity, and the inverse function is not well defined there.

IR_{i_1, \dots, i_r} is considered as a map from the root space to itself, sending $(\lambda_{i_1} \cdots \lambda_{i_r})$ to their inverse. For it to act on the coefficient space, we need to solve the equation $\boldsymbol{\theta}(L) = 0$, take the inverse of the $\lambda_{i_1} \cdots \lambda_{i_r}$ then reconstruct the coefficients. In effect it is $\mathbf{v} \circ IR \circ \mathbf{v}^{-1}$. The chain rule and the implicit function theorem gives

$$J[IR_{i_1, \dots, i_r}(\boldsymbol{\theta})] = J_{\mathbf{v}}|_{\mathbf{v}^{-1}(\boldsymbol{\theta}_{IR})} \text{diag}(1, \dots, -\lambda_{i_1}^{-2}, \dots, -\lambda_{i_r}^{-2} \cdots, 1) J_{\mathbf{v}}^{-1}|_{\boldsymbol{\theta}}$$

We note that we need to pick $S = \{\lambda_{i_1} \cdots \lambda_{i_r}\}$ so that if λ_i is in S then $\bar{\lambda}_i$ is also in S . In that case J_{IR} is real, as it is the Jacobian of a real map, even if $J_{\mathbf{v}}$ could be complex.

Finally by the chain rule and invariance of $\nabla \mathcal{L}(\boldsymbol{\theta})$ under the action of IR gives us the equation for the gradient.

Note if $f(\boldsymbol{\theta}_{IR})$ transforms as

$$f(\boldsymbol{\theta}_{IR}) = h(\boldsymbol{\theta})f(\boldsymbol{\theta})$$

where h is a scalar function and f is a vector function then we have

$$\nabla f(\boldsymbol{\theta}) = \frac{1}{h(\boldsymbol{\theta})} \nabla f(\boldsymbol{\theta}_{IR}) J[IR] - \frac{1}{h(\boldsymbol{\theta})^2} (\nabla h)|_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{IR})$$

If h is given in term of λ , for example $g(\lambda) = \lambda_i^2$ then ∇h could be computed using the Jacobian of the Vieta map

$$\nabla_{\theta} h = \nabla_{g_{\lambda}} J_{\mathbf{v}}^{-1}$$

and from here we can also compute $\nabla f(\theta)$. In practice we only need to compute the gradient of $\bar{\mathcal{L}}$, but it is useful for sanity check to compute gradient of the intermediate terms.

With these relations, we can compute both the value and the gradient of the likelihood function at a non-invertible θ by transforming it to an invertible point where the calculation is numerically stable. Hence we can apply gradient optimization method without any restriction on the domain of θ . Note that we will work with θ that has no multiple roots here where $J[IR]$ is defined.

The root inversion maps have some interesting property near it fixed points as seen in the next lemma.

Lemma 3 *If θ is fixed under a root inversion map $IR = IR_{i_1 \dots i_r}$ then*

$$J[IR]^2(\theta) = I_q \quad (47)$$

In that case, $J[IR]$ has eigenvalues of -1 or 1 only. More over we have

$$\nabla \bar{\mathcal{L}}(\theta) J[IR] = \nabla \bar{\mathcal{L}}(\theta) \quad (48)$$

The first statement is a consequence of the fact that $IR \circ IR = id$ around θ . The second is clear from invariance of the action of IR on the likelihood function and the fact that θ is a fixed point.

We see that this puts constraints on $\nabla \bar{\mathcal{L}}$. If we split the tangent space of θ at a fixed point of IR to eigenspaces of $J[IR]$, corresponding to eigenvalues ± 1 there is no constraint on the eigenspace corresponding to 1 , while if c is an eigenvector corresponding to -1 then $\nabla \bar{\mathcal{L}} \cdot c = 0$. For MA(1) this is already well-known, as $J[IR] = -1$ in that case. It is surprising to us that we can do quite a bit better by examining the eigenvalues of the Jacobian in details. In the paper (Nguyen 2016) we prove that the dimension of the eigenspace corresponding to -1 is

$$\text{mult}_{-1} J[IR](\theta) = \begin{cases} \lfloor r/2 \rfloor + 1 & \text{if } \psi_r = -1 \text{ or } r \text{ is odd} \\ \lfloor r/2 \rfloor & \text{otherwise} \end{cases} \quad (49)$$

Here, $\psi_r = (-1)^r \lambda_{i_1} \cdots \lambda_{i_r}$ and $\lfloor x \rfloor$ denote the integer part of x . From here, we see the only cases where $J[IR] - I_q$ is invertible are $q = 1$,

$\theta = 1 \pm L$ and $q = 2$, $\theta = 1 - L^2$. Those cases are the cases where the constraint are strongest, the corresponding models are critical points of the likelihood function regardless of the sample data set. This is the pile-up effect. In optimization when we observe a critical point close to these values, additional analysis would be required. On the other hand (Davis and Dunsmuir 1996) has studied the both local and global maximum of the likelihood function in detail for MA(1) case. Testing for MA unit root has attracted the attention of several authors, see (Anderson and Takemura 1986; Tanaka 1990; Davis and Dunsmuir 1996; Davis and Song 2011) for the pure MA case. In the later works for MA(1) case, a change of parameter of form $\theta_1 = 1 - \beta/T$ expresses the likelihood function as a function of β , which could have more than one local maximum point. Tests for MA unit roots could be derived from that study. The analysis make use of a join distribution of the Gradient and Hessian with respect to the changed variable β .

Our analysis suggests that when q is larger, the MA unit root constraints are not as strong as when q is small. In the generic cases (corresponding to the hyperplane boundary) where we have one or two conjugated unit roots, we have only one unit root we have at most one constraint on gradient of the likelihood function. At the more complex boundary point the number of constraints is around half the number of unit roots. The constraints could be given very explicitly in term of the unit roots, as we will see in the paper (Nguyen 2016).

We hope the results here provide some help in analyzing unit roots in general case. This topic requires further studies.

7 Invertibility region and initial values

Many results in this section is well-known in the system and control literature. We recall them here for the reader's convenience.

Theorem 4 *The set $\theta_1, \dots, \theta_q$ such that the equation:*

$$1 + \theta_1 L + \dots + \theta_q L^q = 0$$

have roots outside of the unit disc or equivalently the equation:

$$z^q + \theta_1 z^{q-1} + \dots + \theta_q = 0$$

have roots inside the unit disc is a convex, connected set bounded by real, algebraic hyperplanes given by the Schur-Cohn polynomial inequalities.

We refer the readers to the literature (Krein and Naimark 1981; Schur 1917; Cohn 1922; Jury and Anderson 1981; Bistritz 2002) for this classical result and improvements. We do not need the Schur-Cohn boundary explicitly, as it is not too expensive to calculate the roots directly and compare the modulus with one. For readers who are not interested in the details, it is sufficient to know that there exist inequalities formed by algebraic polynomials called the Schur-Cohn polynomials such that the stability restriction on roots are satisfied if and only if these inequalities are satisfied. The Schur-Cohn polynomials could be computed recursively via efficient algorithms in the above references. We will show only a few examples for $q \leq 3$ to illustrate the idea. We note for one variable the condition is simply $-1 \leq \theta_1 \leq 1$, for two variables the condition is

$$\theta_2 < 1$$

$$-\theta_1 + \theta_2 + 1 \geq 0$$

$$\theta_1 + \theta_2 + 1 \geq 0$$

which form a triangle with (inverse) base $\theta_2 = 1$ and top at $(0, -1)$. For three variables the Schur-Cohn conditions are

$$1 + \theta_1 + \theta_2 + \theta_3 > 0$$

$$3 + \theta_1 - \theta_2 - 3\theta_3 > 0$$

$$1 - \theta_1 + \theta_2 - \theta_3 > 0$$

$$1 - \theta_2 - \theta_3^2 + \theta_1\theta_3 > 0$$

The first three conditions give a tetrahedral with vertices $(-1, 3, 3)$, $(1, -1, -1)$, $(1, 3, 3)$, $(-1, -1, 1)$. The last equation restricts it further to a convex region of the tetrahedral. We note the second condition does not appear in the limit determinant of Σ_T^{-1} discussed above. Our simulation shows it is in fact not needed, it seems to be a consequence of the remaining three conditions. We note in our previous discussion of the determinant of Σ_T^{-1} , the factor $\prod(1 - \lambda_i\lambda_j)$ is symmetric and could be expressed as a polynomial in θ_i 's. This function vanishes whenever we have conjugated unit roots so should be closely related

to the invertibility boundary. Up to $q = 3$ this seems to be the only non linear condition. For $q = 4$ that factor is of degree 12 in λ_i , while the Schur-Cohn polynomials are of degree at most 6, so the picture is more complex here. It would be nice to understand more clearly the relationship between the Szegő determinant limits and the Schur-Cohn boundary.

While the invertibility region is convex, in general the likelihood function is not, therefore we need to deal with local minima. Here, the cost function is minus the log-likelihood. We have briefly discuss the situation with root on the unit circle in the previous section so in this section we will focus on optimization technique inside the region. While more theoretical work will be needed to understand the distribution of local minima, our first attempt is to use local optimizers with initial points starting in different sub regions, with the hope that when the mesh of sub regions is fine enough we will catch the global optimum point.

Of course there are many ways to choose the starting points, we describe here the method that we use in our code. Recall that a real polynomial of odd degree always have at least one real root, and in general complex roots always appear in conjugated pairs. We look at inverse of roots of θ , which are roots of $L^q\theta(L^{-1})$. We are assuming they are inside the unit disc. We will call them roots here when there is no confusion.

Our strategy is for real roots, divide the interval $[-1, 1]$ in to regions, and for complex root divide the upper unit disc in to regions, then consider possible arrangements of the q roots to these regions.

To illustrate, let us divide the interval $[-1, 1]$ to three subintervals: $R_1 = (-1, -3^{-1/2}]$, $R_2 = [-3^{-1/2}, 3^{-1/2}]$, $R_3[3^{-1/2}, 1)$. We divide the upper half disc to three regions: C_1 is the half disc with radius $3^{-1/2}$, C_2 is the part of the first quadrant with radius between $3^{-1/2}$ and 1, and C_3 is the part of the second quadrant with radius between $3^{-1/2}$ and 1. The choice of $3^{-1/2}$ is so that the three complex regions to have the same area, and there is exact overlap between the real and complex regions. We could modify the choices some other ways.

Set $q = q_r + 2q_c$, where q_r is the number of real roots and $2q_c$ is the number of complex roots. Consider the arrangements of the q_r real roots to $q_r = q_{r_1} + q_{r_2} + q_{r_3}$ corresponding to the three interval R_1, R_2, R_3 and the $q_c = q_{c_1} + q_{c_2} + q_{c_3}$ complex roots in the upper half plane to the three area C_1, C_2, C_3 with q_{c_1} roots in C_1 , q_{c_2} roots in C_2 and q_{c_3} roots in C_3 .

We can see the number of choices is $\frac{(q_r+1)(q_r+2)}{2}$ for the real roots, and $\frac{(q_c+1)(q_c+2)}{2}$ for the complex roots. So the number of regions under this partition is

$$\sum_{q_c \leq \text{floor}(q/2); q_r = q - 2q_c} \frac{(q_r + 1)(q_r + 2)}{2} \frac{(q_c + 1)(q_c + 2)}{2}$$

It turns out the sum could be simplified to polynomials of degree five depending on q odd or even:

$$\text{number of regions} = \begin{cases} \frac{(q+2)(q+4)(q+6)(q+8)(2q+5)}{1920} & \text{if } q \text{ is even} \\ \frac{(q+1)(q+3)(q+5)(q+7)(2q+13)}{1920} & \text{if } q \text{ is odd} \end{cases} \quad (50)$$

Start with one region, for example we choose say $q_r = q+0+0$ roots on in R_1 and no complex root ($q_c = 0$). The roots could be picked randomly or deterministically. For example we will choose them to be just the middle point of R_1 . Then we construct θ from roots by the Vieta formula.

The resulting θ will be an initial value for the first local optimization. We repeat this for all regions to choose initial points. While the number of initial points grows polynomially, with our algorithm we can compute the likelihood functions relatively fast for practical data size. In practice we pick the best initial points and optimize them further with a local optimizer.

8 Additional topics

8.1 Seasonality and Integration

First we note the whole process work if we add additional drift terms, or additional regressions. For example to allow a polynomial drift we add vectors of form i^k instead of 1 in the definition of \mathbf{X}_{LAG} . Seasonality could be accounted for by seasonal dummy variables, just like the VAR case. We will next discuss integrated models. Consider the following model with scalar θ :

$$\Phi(L)X = \theta(L)\epsilon$$

We note the polynomial division algorithm works for any matrix polynomial and a scalar polynomial. In particular, apply polynomial division of Φ to $L(L-1)$, note that the remainder matrix is a matrix

polynomial of degree at most one we have

$$\Phi(L) = L(L-1)\Gamma(L)t + \Phi_b(1-L) - \Pi L$$

($\Phi_b(1-L) - \Pi L$ is the remainder of the division by $L(L-1)$ which is of degree 1 so will be of form $A + BL$, and we set $\Phi_b = A$, $\Pi = -A - B$).

Let $L = 0$ and $L = 1$, respectively we get:

$$\Phi_b = I_k$$

$$\Pi = -\Phi_L(1) = -I_k + \Phi_1 + \dots + \Phi_p$$

Let $\Delta = 1 - L$. The equation becomes:

$$\Delta \mathbf{X}(t) = \Gamma(L)\Delta \mathbf{X}(t-1) + \Pi \mathbf{X}(t-1) + \boldsymbol{\theta}(L)\epsilon(t)$$

Apply $\boldsymbol{\theta}(L)^{-1}$ to both sides we get

$$\Delta \mathbf{X}_{\theta,t} = \Gamma(L)\Delta L \mathbf{X}_{\theta,t} + \Pi L \mathbf{X}_{\theta,t} + \epsilon_t$$

where $\mathbf{X}_{\theta,t}$ is $\boldsymbol{\theta}(L)^{-1}\mathbf{X}(t)$. This is our VECM form. We can apply an argument similar to Johansen for cointegration here. We construct a regression between $\Delta \Theta_T^{-1} \mathbf{X}_t$ and the lags represented by $\mathbf{X}_{\theta, \text{LAG}}$, where $\mathbf{X}_{\theta, \text{LAG}}$ consists of terms $\Delta \Theta_T^{-1} L^1 \mathbf{X}_{t-1}, \dots, \Delta \Theta_T^{-1} L^p \mathbf{X}_{t-p+1}$ and $\Theta_T^{-1} L \mathbf{X}_t$. This is essentially the same construction of the VARMA case, the integration component correspond to the term $\Theta_T^{-1} L \mathbf{X}$. Let $r \leq k$ be the rank of Π . If the $r = k$ then we have a stationary process. If $r = 0$ we do not have cointegration. If $0 < r < k$ then we have a cointegrating system. We can decompose $\Pi = \alpha \beta^T$ with α, β are a $k \times r$ matrix of full rank. It remains to apply a rank test to figure out the rank r . We expect a result similar to Johansen's test (Johansen 1991) where the inner product defined by Σ_T plays a role.

8.2 Extension to infinite component MA

Next, a few words about the case when we have an infinite number of MA components. If we aim to study models with a finite number of VAR terms but an infinite MA scalar terms, we expect the result here to carry through, provided we apply the appropriate inner product constructed from the MA scalar terms. So the issue is to study this inner product. The survey paper (Bingham 2012) provided a framework to think about the MA(∞) case. Blaschke product used by Hansen and Sargent is closely related to Hardy spaces, so it has

been understood for sometime that Toeplitz operators, Wiener-Hopfs, Hardy spaces are what needed to extend the theory to infinite component moving average models. In a future paper we hope to work out the technical details.

Since we deal with an infinite past, a rigorous approach may require more analytic machinery than we intent to cover here, but let us sketch a few ideas. When we have infinite MA terms, the invertible condition is just the condition that $\theta(L)$ has no root or pole inside or on the unit circle (see the next section for the case of poles on the unit circle - as in case of fractional Gaussian). θ is called an *outer function* or *Szegő function*. An example of such function could be any stable ARMA rational function (we presumably formulate that all entries of $\Phi(L)$ has a polynomial factor f and use f/θ as our MA(∞) function). A function of form $(1-b_1L)^{\alpha_1} \dots (1-b_mL)^{\alpha_m}$ with $|b| < 1$ is also an outer function. We expect to be able to apply our framework to calibrate a finite number of parameters that generate a model with infinite moving average components.

Σ_T and K are finite dimension but now \bar{K}_T is of infinite dimension. We will need a definition of Gaussian measure as well as determinant in this context - both of which fortunately have been studied for a long time. The discussion of infinite dimensional MA with analytic outer function would hopefully provide error estimates to our main regression of Φ when we cut off the expansion of θ by a finite number $q_c(\theta)$ of terms.

Let us shift the index by 1 and consider the index set of the sample as $\{0, \dots, T-1\}$ instead of $\{1, \dots, T\}$. This makes it more convenient when we write convolutions. Set $\theta_+(L) = \theta(L)$ and $\theta_-(L) = \theta(L^{-1})$, considered as Laurent series. Consider the vector space $V_{\geq 0}$ spanned by basis $\{v_i\}_{i=0}^{\infty}$. For any Laurent series $a(L) = \sum_{-\infty}^{\infty} a_i L^i$ define the infinite Toeplitz matrix $\mathcal{T}_{\infty}(a) = (a_{j-k})_{j,k=0}^{\infty}$. This is the matrix of the action of convolution of $a(L)$ on $V_{\geq 0}$:

$$a.v_k = \sum_{j=0}^{\infty} a_{j-k} v_j$$

(we will need a norm for the sum to make sense). In our paper we work with the top $T \times T$ block of this matrix. We note $\mathcal{T}_{\infty}(\theta_+)$ is the infinite version of Θ_T , $\mathcal{T}_{\infty}(\theta_-)$ is the infinite version of Θ_T^{-1} , $\mathcal{T}_{\infty}(\theta_+\theta_-)$ is the full autocovariance matrix, and its upper left $T \times T$ matrix is our Σ_T . In their second proof of the Borodin Okounkov's

formula (Basor and H. 2000), the authors defined a matrix A as

$$A = \mathcal{T}_\infty(\boldsymbol{\theta}_+^{-1})\mathcal{T}_\infty(\boldsymbol{\theta}_+\boldsymbol{\theta}_-)\mathcal{T}_\infty(\boldsymbol{\theta}_-)$$

and showed

$$A^{-1} = \mathcal{T}_\infty(\boldsymbol{\theta}_-\boldsymbol{\theta}_+)\mathcal{T}_\infty(\boldsymbol{\theta}_+\boldsymbol{\theta}_-)$$

we note $K(\boldsymbol{\theta}, T)$ is just the upper left $T \times T$ block of A^{-1} . They noted that $A^{-1} - I$ is of trace class. We have shown earlier in case $\boldsymbol{\theta}(L)$ is polynomial this trace class part is $\lambda\lambda'$. To define λ in the MA(∞) term case we will need some analysis tools which we will not get in to in this paper but formally we can mimic the definition of the polynomial case and define it as an infinite dimensional matrix. We note Θ_* now acts on an infinite dimensional Hilbert space corresponding to $\epsilon_{i < 0}$ (note we shifted the indices by 1), Θ_T is defined as before and $\lambda = \Theta_T^{-1}\Theta_*$ is a linear operator represented by a matrix with columns indexed by negative integers and row indexed by $\{0, \dots, T-1\}$. The interested reader could work out the AR(1) case where $\boldsymbol{\theta}(L) = \sum_{i=0}^{\infty} \phi^i L^i$ and find $\lambda = (\lambda_{ij})_{i=0, j=-\infty}^{i=\infty, j=-1}$ with $\lambda_{0j} = \phi^{-j}$ and $\lambda_{ij} = 0$ with $i \neq 0$. From here $(\lambda\lambda')_{ij} = \frac{\phi^2}{1-\phi^2}$ if $i = j = 0$ and zero otherwise, and get to the exact likelihood function of AR(1).

We see the determinant of Σ_∞ could now be expressed in two different ways, $\det(I + \lambda\lambda') = \det(I + \lambda'\lambda)$. (Basor and H. 2000) showed the first determinant is the same Fredholm operator determinant in Borodin-Okounkov's formula. In either AR or MA case we expect one of the determinants to collapse to a finite dimensional determinant, but in general we have two Fredholm operator determinant expressions of $\det \Sigma_T$. We note that if the coefficients decay sufficiently after $q_c < T$ terms, we only need q_c MA terms in the second expression. We note although we have an infinite (or q_c) number of MA terms, in general they are controlled by a finite (and smaller than q_c) number of parameters and the gradient calculation would also apply with appropriate application of the chain rule. We expect our calibration method would still be effective in this last case, however to be practical the models need to be in special forms for us to check the invertibility condition.

8.3 Fractional VARMA

We again assume finite dimensional VAR model, with a fractional Gaussian MA component

$$(1 + \Phi_1 L + \cdots \Phi_p L^p)(1 - L)^d X_t = (1 + \theta_1 L + \cdots \theta_q L^q) \epsilon_t$$

Here the scalar function to consider is

$$\theta_d(L) = (1 - L)^{-d} \theta(L)$$

which could be written in $\text{MA}(\infty)$ form. We conjecture the main theorem is still valid in the form given by the matrix Σ_T which is finite dimensional, however careful analysis is needed to define λ , as seen in the previous section. We note the determinant Σ_T tends to infinitive at large T . If we apply mechanically the Szegö limit theorem we see beside the inverse polynomial terms, the determinant $\det(\Sigma_T)$ would have an extra term corresponding to $d \log(1 - L)$:

$$\exp\left(\sum_{k=1}^T \frac{d^2}{k}\right)$$

which increases as T^{d^2} . This is a special case of the Fisher-Hartwig conjecture (Fisher and Hartwig 1969) which has been proved for some time (Ehrhardt 2001). In fact in Toeplitz operator literature, people consider function with several (conjugated) poles on the unit circle, as well as other types of singularities. The analysis near singular/zero points on the unit circle would need careful analysis, and we hope operator theory method to be helpful here.

We note that while there need to be theoretical justifications, invertibility considerations and initial point selection, for the last few sections, the algorithms and coding require little modifications. As these models are dependend on a finite set of parameters p_i , we only need functions to supply the coefficients θ_i and the gradient matrix $\frac{\partial \theta_i}{\partial p_j}$, which will be model dependent.

9 Conclusion

We have tested the likelihood function and calibration algorithm presented here in R and C++ codes. The philosophy of replacing the scalar MA components with an inner product defined by the finite

Toeplitz matrix seems fruitful and we expect many other results related to Vector Auto Regressive models are to have corresponding VARMA analogues. It remains to be seen how the calibration algorithm suggested here applies in practical forecast.

Appendix A A few matrix facts

Lemma A.1 *Let X, Y, β, K, Ω be matrices with compatible dimension such that the following expression is well formed*

$$\text{Tr}((Y' - \beta' X')K(Y - X\beta)\Omega) \quad (51)$$

Assume further, that K and Ω are invertible symmetric positive definite matrices. Also assume $(X'KX)$ is invertible. With X, Y, K, Ω known, the above expression has its minimum at

$$\beta_{opt} = (X'KX)^{-1}X'KY$$

and thus β is independent of Ω .

Proof. Set $\beta = \beta_{opt} + b$ and expand the expression.

$$\begin{aligned} \text{Tr}((Y' - \beta' X')K(Y - X\beta)\Omega) &= \text{Tr}((Y' - \beta'_{opt} X')K(Y - X\beta_{opt})\Omega) - \\ &\text{Tr}(b' X' K(Y - X\beta_{opt})\Omega) - \text{Tr}((Y' - \beta'_{opt} X')KXb\Omega) + \\ &\text{Tr}((b' X' KXb)\Omega) \end{aligned} \quad (52)$$

Now note

$$\begin{aligned} (X'KX)\beta_{opt} &= X'KY \\ \beta'_{opt}(X'KX) &= Y'KX \end{aligned}$$

We see both middle terms are zero, while the first and last terms are positive because Kronecker product of positive definite matrix K and Ω is also positive definite and applying lemma 2. So the minimum is attained at $b = 0$.

This proves the optimality of $\begin{pmatrix} \mu \\ \Phi \end{pmatrix}_{opt}$.

The following is already well-known:

Lemma A.2 *Assuming K is positive definite. Then*

$$P_K = K - KM(M'KM)^{-1}M'K \quad (53)$$

is positive semi-definite for all M such that $M'KM$ is invertible.

Proof: Consider a decomposition $K = L'L$ and set $LM = M_1$. We see

$$L'^{-1}P_K L^{-1} = I_T - LM(M'L'LM)^{-1}M'L' = I_T - M_1(M_1'M_1)^{-1}M_1'$$

is a projection, and hence has eigenvalues 0 and 1. So P_K is positive semi-definite.

So we have in particular Ω_{opt} is positive semi-definite, regardless of the sample data \mathbf{X} . In practice there may exist data \mathbf{X} such that Ω has a zero-eigenvalue. Some regularization need to fix K for that case.

Appendix B Simulation results

Using our R script we have tested and confirmed the relationship between Σ_T and K and \bar{K} . We also have confirmed the Szegö limit of the determinant. We also have confirmed the invariant of the likelihood function under the root inversion algorithm, this was done against small sample as large sample data would lead to implosion in intermediate steps.

Using simulated data then maximizing the likelihood function we are also able to recover original models in our test cases. This includes

- simple $p = 0, q = 2$ models (with 2×2 scalar θ).
- ARMA $p = 2, q = 2$ model
- VARMA models with $k = 2, (p = 1, q = 1)$ matrix polynomials:

$$\Phi_1 = - \begin{pmatrix} 0.02284288 & 0.4027705 \\ 1.06073525 & -0.2589487 \end{pmatrix}$$

$$\Theta_1 = - \begin{pmatrix} -0.4100472 & 0.3227580 \\ 2.1013041 & 0.2378265 \end{pmatrix}$$

and $X_t = (I - \Phi_1 L)X_t + (I + \Theta_1 L)\epsilon_t$. We write the respective matrix polynomials $\Theta^M(L)$ and $\Phi^M(L)$. (The long decimals in the matrices were due to the fact we ran a simulation to search for stable matrices.) This is equivalent to a $p = 2, q = 2$ model with scalar $\theta = \det(\Theta^M(L))$, and the AR term $\text{adj}(\Theta^M(L)\Phi^M(L))$. We are able to recover both the scalar denominator and the degree 2 matrix polynomial numerator.

- VARMA model with $k = 2$ given by $(p = 2, q = 2)$ matrix polynomials. This is equivalent to $(p = 4, q = 4)$ scalar MA model. Again we recovered the equivalent scalar-denominator model.
- VARMA model with $k = 4$ with $p = 5, q = 3$ where the numerator is a matrix polynomial and the denominator is a scalar polynomial. While in the previous two cases, we were able to find the optimal parameters by an optimization with initial vector at 0, for the last case we had to apply the partition of the invertibility region mentioned above.

ACKNOWLEDGEMENT. We would like to thank Thong Nguyen for very helpful suggestions and help with literature. He filled our gap in knowledge in time series and statistics through insightful conversation and providing reading material. He pointed out (16) is essentially the Yule-Walker equations, with a different twist. We thank Utkarsh Samant for encouragement and providing infrastructure where much of test was carried out. Any error remains our responsibility alone.

Bibibliography

- Anderson, T. W. and A. Takemura (1986). “Why do noninvertible estimated moving averages occur?” In: *Journal of Time Series Analysis* 7, 235–254.
- Aoki, Masanao (1987). *State Space Modeling of Time Series*. Universitext. Springer.
- Basor, E.L. and Widom H. (2000). “On a Toeplitz determinant identity of Borodin and Okounkov”. In: *Integral Equations and Operator Theory* 37, pp. 397–401.
- Bingham, N.H. (2012). “Szeg’s theorem and its probabilistic descendants”. In: *Probab. Surveys* 9, pp. 287–324. DOI: 10.1214/11-PS178. URL: <http://dx.doi.org/10.1214/11-PS178>.
- Bistritz, Y. (Mar. 2002). “Zero location of polynomials with respect to the unit-circle unhampered by nonessential singularities”. In: *IEEE Trans. on Circuits and Systems, part I* 49, pp. 305–314.
- Borodin, A. M. and A. Okounkov (2000). “A Fredholm determinant formula for Toeplitz determinants”. In: *Integral Equations and Operator Theory* 37, pp. 386–396.

- Box, George and Gwilym Jenkins (1970). *Time Series Analysis: forecasting and control*. Holden-Day Series in Time Series Analysis and Digital Processing. Holden-Day Inc.
- Cohn, A. (1922). “Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise”. In: *Math. Zeit.* 14, pp. 110–148.
- Davis, R. A. and William T. M. Dunsmuir (1996). “Maximum Likelihood Estimation for MA(1) Processes with a Root on or near the Unit Circle”. In: *Econometric Theory* 12.1, pp. 1–29. ISSN: 02664666, 14694360. URL: <http://www.jstor.org/stable/3532753>.
- Davis, Richard A. and Li Song (Dec. 2011). “Unit roots in moving averages beyond first order”. In: *Ann. Statist.* 39.6, pp. 3062–3091. DOI: 10.1214/11-AOS935. URL: <http://dx.doi.org/10.1214/11-AOS935>.
- Ehrhardt, T. (2001). “A status report on the asymptotic behavior of Toeplitz determinants with Fisher-Hartwig singularities”. In: *Recent Advances in Operator Theory (Groningen, 1998)*. Ed. by K. E. Shuler. Vol. 124. Oper. Theory Adv. Appl. Basel: Birkhäuser.
- Fisher, M. E. and R. E. Hartwig (1969). “Toeplitz Determinants: Some Applications, Theorems, and Conjectures”. In: *Advances in Chemical Physics: Stochastic Processes in Chemical Physics*. Ed. by K. E. Shuler. Vol. 15. Hoboken, NJ, USA.: John Wiley & Sons, Inc. Chap. 18.
- Geronimo, J. S. and K. M. Case (1979). “Scattering theory and polynomials orthogonal on the unit circle”. In: *J. Math. Phys.* 20, pp. 299–310.
- Gohberg, I., M.A Kaashoek, and F van Schagen (1987). “Szeg-Kac-Achiezer formulas in terms of realizations of the symbol”. In: *Journal of Functional Analysis* 74.1, pp. 24–51. ISSN: 0022-1236. DOI: [http://dx.doi.org/10.1016/0022-1236\(87\)90037-1](http://dx.doi.org/10.1016/0022-1236(87)90037-1). URL: <http://www.sciencedirect.com/science/article/pii/0022123687900371>.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.
- Hansen, Lars Peter and Thomas J. Sargent (1980). “Formulating and estimating dynamic linear rational expectations models”. In: *Journal of Economic Dynamics and Control* 2, pp. 7–46. ISSN: 0165-1889. DOI: [http://dx.doi.org/10.1016/0165-1889\(80\)90049-4](http://dx.doi.org/10.1016/0165-1889(80)90049-4). URL: <http://www.sciencedirect.com/science/article/pii/0165188980900494>.

- Harvey, David (Jan. 2011). “Faster algorithms for the square root and reciprocal of power series”. In: *Mathematics of Computation* 80.273, 387394.
- Johansen, Sren (1991). “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models”. In: *Econometrica* 59.6, pp. 1551–1580. ISSN: 00129682, 14680262. URL: <http://www.jstor.org/stable/2938278>.
- Jury, E. and B Anderson (1981). “A note on the reduced Schur-Cohn criterion”. In: *IEEE Transactions on Automatic Control* 26.2, pp. 612–614. DOI: 10.1109/TAC.1981.1102662.
- Klein, André and Guy Mélard (Apr. 2014). “An algorithm for the exact Fisher information matrix of vector ARMAX time series”. In: *Linear Algebra and its Applications* 446, 124.
- Kramer, M. and M. Rosenblatt (July 1993). “The Gaussian log likelihood and stationary sequences”. In: *Developments in Time Series Analysis*. Ed. by Rao T. S. London: Chapman & Hall, pp. 69–79.
- Krein, M. G. and M. A. Naimark (1981). *The method of symmetric and Hermitian forms in the theory of the separation of roots of algebraic equations*. Vol. 10. Linear and Multilinear Algebra. Originally in Russian, Kharkov 1936. Springer, pp. 265–308.
- Nguyen, Du (2016). “Jordan Decomposition of Jacobian of the Root Inversion maps with Application to Moving-Average Unit Roots”. In: in preparation.
- Pnevmatikakis, E. A et al. (2014). “Fast Kalman Filtering and Forward-Backward Smoothing via a Low-Rank Perturbative Approach”. In: *Journal of Computational and Graphical Statistics* 23.2, pp. 316–339.
- Schur, I. (1917). “Über Potenzreihen, die in Innern des Einheitskreises Beschränkt Sind”. In: *Journal für die Reine und Angewandte Mathematik* 147. and vol. 148, pp. 122–145, Berlin, 1918., pp. 205–232.
- Szegő, G. (1952). “On certain Hermitian forms associated with the Fourier series of a positive function”. In: *Comm. Sm. Math. Univ. Lund [Medd. Lunds Univ. Mat. Sem.* 228238.

- Tanaka, Katsuto (Dec. 1990). “Testing for a Moving Average Unit Root”. In: *Econometric Theory* 6 (04), pp. 433–444. ISSN: 1469-4360. DOI: 10.1017/S0266466600005442. URL: <http://journals.cambridge.org/a>
- Tsay, Ruey (1991). “Two Canonical Forms for Vector ARMA processes”. In: *Statistica Sinica* 1, pp. 247–269.